BMC Medical Research Methodology

BioMed Central

Research article

Open Access

Randomized trials, generalizability, and meta-analysis: Graphical insights for binary outcomes

Stuart G Baker*[1] and Barnett S Kramer[2]

Address: [1]Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, USA and [2]Office of Disease Prevention, National Institutes of Health, USA

Email: Stuart G Baker* - sb16i@nih.gov; Barnett S Kramer - KramerB@OD.NIH.GOV

* Corresponding author

Published: 16 June 2003

BMC Medical Research Methodology 2003, **3**:10

Received: 28 March 2003

Accepted: 16 June 2003

This article is available from: http://www.biomedcentral.com/1471-2288/3/10

Abstract

Background: Randomized trials stochastically answer the question. "What would be the effect of treatment on outcome if one turned back the clock and switched treatments in the given population?" Generalizations to other subjects are reliable only if the particular trial is performed on a random sample of the target population. By considering an unobserved binary variable, we graphically investigate how randomized trials can also stochastically answer the question, "What would be the effect of treatment on outcome in a population with a possibly different distribution of an unobserved binary baseline variable that does not interact with treatment in its effect on outcome?"

Method: For three different outcome measures, absolute difference (*DIF*), relative risk (*RR*), and odds ratio (*OR*), we constructed a modified BK-Plot under the assumption that treatment has the same effect on outcome if either all or no subjects had a given level of the unobserved binary variable. (A BK-Plot shows the effect of an unobserved binary covariate on a binary outcome in two treatment groups; it was originally developed to explain Simpsons's paradox.)

Results: For *DIF* and *RR*, but not *OR*, the BK-Plot shows that the estimated treatment effect is invariant to the fraction of subjects with an unobserved binary variable at a given level.

Conclusion: The BK-Plot provides a simple method to understand generalizability in randomized trials. Meta-analyses of randomized trials with a binary outcome that are based on *DIF* or *RR*, but not *OR*, will avoid bias from an unobserved covariate that does not interact with treatment in its effect on outcome.

Background

Consider a randomized trial in which subjects are randomized to either a control or experimental intervention. The approach to statistical inference depends on the question one would like to answer.

One question is "What would be the effect of an intervention on outcome if we turned the clock backwards so that subjects randomized to the experimental treatment received the control treatment and vice versa?" Of course this question cannot be answered empirically by direct observation because one cannot go back in time. In a landmark paper on causal inference, Rubin [1] presented a stochastic answer, demonstrating that the estimated treatment effect in a randomized trial is an unbiased *estimate* of the treatment effect if the clock were turned

backwards and the treatments were reversed. Rubin [1] noted that estimates are generalizable to a target population if the subjects in the study are a random sample from the target population. (See [2] and [3] for additional discussions of the Rubin causal model including the requirement that the effect of treatment on one subject is independent of the effect of treatment on another subject.)

A broader question is "What is the effect of intervention in a different population that is not a random sample from the target population?" This question cannot be answered empirically. (In fact, if it were required for valid generalization of results, it would present a serious limitation of the scientific method in medical decision making.) In the most general situation in which the treatment effect varies by population, the question is also unanswerable stochastically. However a restricted version of this question can be answered stochastistically. Our starting point is to postulate an unobserved baseline binary random variable. Unobserved baseline variables have often been considered in discussing randomization. According to Meier [4] "...the role of randomization is to distribute the effects of baseline variables, both measured ones and those not observed, in such a way that the statistical analysis makes due allowance for them. It is precisely when there are hidden variables which may be influential that randomization is most important." To make progress we assume no interactive effect on probability of outcome between the unobserved binary variable and treatment. This assumption lies at the core of our ability to generalize results of clinical trials to populations other than those from whom the original sample in the trial was drawn. For some hypothetical situations where the non-interaction assumption for an unmeasured variable would be violated, see [5].

Using the above framework, we address the following question, "What is the effect of intervention in a population in which a different fraction have an unobserved binary variable that does not interact with treatment in its effect on outcome?" We investigate this question for three common outcome measures, absolute difference (*DIF*), relative risk (*RR*), and odds ratio (*OR*).

In related work, Gail et al [5] estimated the bias when one fits a model without an unobserved variable to data generated from a randomized trial with an unobserved variable that does not interact with treatment in its effect on outcome. For binary outcomes they found no bias with *DIF* and *RR* but a bias with *OR*. However their complex formulas provide little insight to the general health professional and do not directly address our question related to generalizability. In other related work, Anderson et al [6] also showed no bias with linear and exponential (i.e. multiplicative) models in the presence of an unobserved variable. Although Anderson et al [6] presented a plot, related to the BK-Plot, showing the effect of a continuous unobserved variable, they did not relate the plot to generalizabilty.

Methods

We start with a standard BK-Plot (Figure 1, left side) based on hypothetical scenarios. The BK-Plot was originally developed as a graphical approach to explain Simpson's Paradox [7,8] and extended to other problems [9]. The horizontal axis is the fraction of subjects with the unobserved baseline variable at a given level. The vertical axis is the probability of outcome, such as treatment success. The plotted lines indicate the probability of outcome as a function of the unobserved binary variable. One line corresponds to subjects randomized to the control group, and the other line corresponds to subjects randomized to the treatment group.

We consider three common outcomes measures: the absolute difference in probability of outcome (*DIF*), the relative risk (*RR*), and the odds ratio (*OR*). Absolute difference is derived from an additive model on the original scale; relative risk is derived from a multiplicative model on the original scale as plotted here (or an additive model on a logarithmic scale); odds ratio can be plotted on the original scale as done here, but is often derived from an additive model on a logistic scale.

For each outcome measure we present a BK-Plot under the assumption of no-interaction between treatment and the two levels of the unobserved binary variable in their effect on the outcome measure. In other words, to fulfill the condition of no interaction between the treatment and the unobserved binary variable, the outcome measure comparing treatment groups, whether *DIF*, *RR* or *OR*, has the same value at the leftmost and rightmost points on the horizontal axis. As the fraction of subjects with a given level of the binary variable varies from 0 to 1, the BK-Plot traces a linear combination of the outcome measure from the leftmost to the rightmost points (Figure 1, left side).

To investigate how the outcome measure changes as the proportion of subjects with a given level of the unobserved binary variable varies from 0 to 1, we present a modified BK Plot (Figure 1, right side), in which the outcome measure is plotted against the fraction with the unobserved binary variable. Because we assumed no interactive effect on the outcome measure between the unobserved binary variable and treatment, the leftmost and rightmost points of the plots on the right side of the Figure are constrained to be equal.

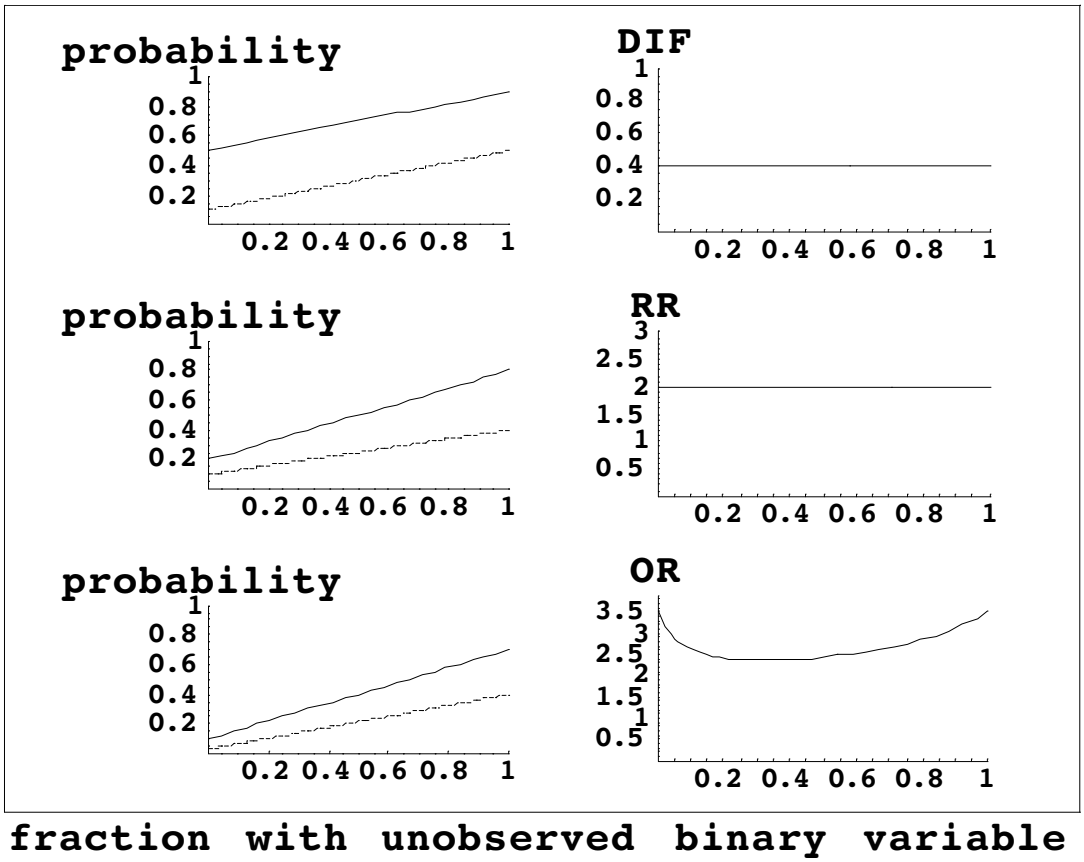

Figure 1
The left side represents a standard BK-Plot, where the diagonal lines correspond to the probabilities of outcome in two randomization groups as a function of the fraction of subjects with the unobserved binary variable. The right side depicts a modified BK-Plot, where the outcome measure is plotted as a function of the fraction of subjects with the unobserved binary variable. We assume no interaction between the unobserved binary variable and treatment effect on the probability of outcome. Graphically, this means that we created BK-Plots so that the outcome measure has the same value at the leftmost and rightmost points. *DIF* = absolute difference; *RR* = relative risk; *OR* = odds ratio.

Results

Based on Figure 1, for *DIF* and *RR*, but not *OR* the outcome measure was constant as the fraction of subjects with a given level varied from 0 to 1. Although the graphic is insightful, for the interested reader we provide the following algebraic derivation of these results. Suppose the randomization groups are labeled z = treatment A or treatment B. Let x = 0 or 1 denote the two levels of the unobserved binary variable. Let p denote the proportion of subjects with the unobserved binary variable at x = 1. Let $g_z(p)$ denote the probability of outcome in randomization group z when a fraction p have the unobserved variable at level x = 1. Let f_{xz} denote the probability of outcome in randomization group z when all subjects are at level x of the unobserved variable. (This represents the rightmost point of the horizontal axis in Figure 1 when x = 1). The marginal probabilities, i.e. the probabilities of outcome when a fraction p have the unobserved variable at level x = 1, are

$$g_A(p) = f_{0A}(1 - p) + f_{1A}\,p$$

$g_B(p) = f_{0B}(1 - p) + f_{1B}\,p$.

For an additive model, the outcome measure is the absolute difference, $f_{xA} - f_{xB}$. Under the assumption of no interaction between treatment effect and the unobserved binary variable, $f_{xA} - f_{xB} = DIF$ for $x = 0, 1$. This implies a constant difference in marginal probabilities, namely $g_A(p) - g_B(p) = DIF$, which holds for all values of p.

For a multiplicative model, the outcome measure is the relative risk, f_{xA}/f_{xB}. Under the assumption of no interaction between treatment effect and the unobserved binary variable, $f_{xA}/f_{xB} = RR$ for $x = 0, 1$. This implies a constant ratio of marginal probabilities, namely, $g_A(p)/g_B(p) - RR$, which holds for all values of p.

The results differ when the outcome measure is the odds ratio, $f_{xA}(1 - f_{xB})/(f_{xB}(1 - f_{xA}))$. Under the assumption of no interaction between treatment effect and the unobserved binary variable, $f_{xA}(1 - f_{xB})/(f_{xB}(1 - f_{xA})) - OR$ for $x = 0, 1$. However, this does *not* imply that $g_A(p)(1 - g_B(p))/(g_B(p)(1 - g_A(p))) = OR$ for all p. In the Appendix we present a calculation to quantify the possible bias from using *OR* in a particular trial.

Discussion

There is a large literature discussing the relative merits of using *RR*, *DIF*, and *OR* as outcome measures [10]-[14]. Our results concerning generalizability of *DIF* and *RR*, but not *OR*, in the presence of an unobserved binary covariate with no interaction, add important new information to this discussion.

Because the analyst must weight all the issues, we think it is helpful to present our perspective on some of the other factors that affect the choice of outcome measure. We believe the outcome measure should reflect the underlying model if it is known. Also we agree that one should consider how well the model of constant *RR*, *DIF*, *OR* fits the data [10].

It is sometimes argued that *DIF* and *RR* should *not* be used because extrapolated estimates might violate the constraints that $0 < DIF < 1$ and $RR > 0$ [10]. (For example, suppose that in 9 trials the probability of outcome in the control group is .1 and the probability of outcome in the intervention group is .6. so $DIF = .5$. Also suppose that in 1 additional trial, the probability of outcome in the control group is .65 and the probability of outcome in the intervention group is .95 so $DIF = .3$. If all trials are equal size, a weighted estimate of *DIF* with weights inversely proportional to the variance yields $DIF_{avg} = .47$. The estimated probability of outcome in the last trial would then be $.65 + DIF_{avg} = 1.12$, which violates the constraint on *DIF*.) In contrast to many other investigators we are not concerned with this extrapolation problem. In many meta-analyses the extrapolated estimates will not violate the constraints. If an extrapolated estimate violates a constraint, it could be a valuable indication that the model is inappropriate when applied to all the studies. If the constraint is violated only slightly, it might be sensible to fit a model that constrains *DIF* and *RR* to lie in valid ranges [11].

Sometimes it is argued that *RR* should not be used because its value changes if the labels of the binary outcome are reversed [10]. In particular, if *RR* is constant with one set of labels it is typically not constant if the labels are reversed. However, because the labels have an important meaning (e.g. survive or die), we are not concerned that *RR* changes with label reversal. In contrast, in latent class models, the class labels are arbitrary, so it is helpful to check the computations by verifying that the results are the same if the labels are reversed. A more serious criticism of *RR* is sensitivity to small counts [12]. We agree with this criticism and do not recommend using *RR* with small counts in one group.

We agree with much of the literature that, in terms of interpretation, *RR* and *DIF* are preferable to *OR*. According to Sackett et al [14] "because very few clinicians are facile at dealing with odds and relative odds, ORs are not useful in their original form at the beside or examining room". Walter [10] writes, "The OR is undeniably the most difficult measure to intuit, so it likely to be less useful than RD [*DIF*] or RR for communicating risk"

Besides the choice of outcome measure, other factors affect the appropriateness of combining results from randomized trials and should be considered by the analyst. One factor is the variation in all-or-none compliance among trials. To reduce the variation from this factor, one can fit a model based on inherent compliance (i.e., with baseline subgroups "always-takers", "compliers", and "never-takers") [15,16]. These models have been applied to meta-analyses involving *DIF* as an outcome [17,18]. Related models for *RR* [19,20] could be used for meta-analyses involving *RR*. Our graphic supporting the use of *DIF* and *RR* would directly apply to "compliers", who are the subgroup of interest in these models for all-or-none compliance.

Another factor affecting the combination of results from randomized trials is the variation in treatment (e.g. variation in doses or levels of ancillary care). Despite the theoretical results in this paper, a large empirical study comparing the use of *RR* and *OR* in meta-analyses found little difference in heterogeneity when using *RR* and *OR* [21]. A likely explanation is that the impact of variations in treatment was larger than the bias from using *OR*.

Conclusion

The issue of generalizability of randomized trials is important in meta-analyses of randomized trials. To avoid bias from an unobserved binary variable that does not interact with treatment in its effect on outcome (and hence increase generalizability of results), one should use *DIF* or *RR*, but not *OR*, as an outcome measure.

Authors' Contributions

SGB wrote the initial draft. BSK made substantial improvements to the manuscript.

Appendix

If one has data from a randomized trial, the following calculation shows the possible bias from using *OR* with no interaction between treatment effect and the unobserved binary variable. Suppose the fraction of subjects with the unobserved binary variable is $p = .5$. From the trial we can estimate $g_A = g_A(.5)$ and $g_B = g_B(.5)$. With $p = .5$, f_{0z} will be the same distance above g_z as f_{1z} is below g_z. Therefore we can write $f_{0A} = g_A(1 - s)$, $f_{1A} = g_A(1 + s)$, $f_{0B} = g_B(1 - k)$, and $f_{1B} = g_B(1 + k)$, where $k \leq \text{minimum}(1/g_B - 1, 1)$ and $s \leq \text{minimum}(1/g_A - 1, 1)$. Let $OR^* = g_A(1 - g_B)/(g_B(1 - g_A))$ denote the apparent odds ratio. Let $\text{OR}^*_x = f_{xA}(1 - f_{xB})/(f_{xB}(1 - f_{xA}))$ denote the true odds ratio when all or none of the subjects have the unobserved covariate. Under the assumption of no interaction between the unobserved covariate and treatment effect, $\text{OR}^*_0 = \text{OR}^*_1$. Solving this equation for s gives

$$s = \frac{1 - g_B + g_B k^2 - \sqrt{[-4g_A k(k - g_A k) + (-1 + g_B - g_B k^2)^2}}{2g_A k}.$$

Substituting the above formula for s into OR^*_0 gives a function of k that we denote $\text{OR}^*_0(k)$. This function represents possible values for the true odds ratio. For example, if $g_A = .2$ and $g_B = .4$, the apparent odds ratio is $OR^* = .375$. However under the model the true odds ratio could have values $\text{OR}^*_0(.3) = .36$, $\text{OR}^*_0(.5) = .32$, or $\text{OR}^*_0(.9) = .20$.

References

1. Rubin DB: **Estimating causal effects of treatments in randomized and nonrandomized studies** *Journal of Educational Psychology* 1974, **66:**688-701.
2. Holland PW: **Statistics and Causal Inference** *Journal of the American Statistical Association* 1986, **81:**945-960. (with discussion) to page
3. Little RJ and Rubin DB: **Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches** *Annual Review of Public Health* 2000, **21:**121-145.
4. Meier P: **Statistics and medical experimentation** *Biometrics* 1975, **31:**511-529.
5. Gail MH, Wieand S and Piantadosi S: **Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates** *Biometrika* 1984, **71:**431-444.
6. Anderson S, Auquier A, Hauck WW, Oakes D, Vandele W and Weisberg HI: **Statistical Methods for Comparative Studies** *Techniques for Bias Reduction John Wiley & Sons, New York*; 1980.
7. Wainer H: **The BK-Plot: Making Simpson's paradox clear to the masses** *Chance* **15:**60-62.
8. Baker SG and Kramer BS: **Good for women, good for men, bad for people: Simpson's paradox and the importance of sex-specific analysis in observational studies** *Journal of Women's Health & Gender-Based Medicine* 2001, **10:**867-872.
9. Baker SG and Kramer BS: **The transitive fallacy for randomized trials: If A bests B and B bests C in separate trials, is A better than C? BMC Medical Research Methodology 2002** *BMC Medical Research Methodology* 2002, **2:**13[http://www.biomedcentral.com/1471-2288/2/13].
10. Walter SD: **Choice of effect measure for epidemiological data** *Journal of Clinical Epidemiology* 2000, **53:**931-939.
11. Warn DE, Thompson SG and Spiegelhalter DJ: **Bayesian random effects meta-analysis of trials with binary outcomes: methods for absolute risk difference and relative risk scales** *Statistics in Medicine* 2002, **21:**1601-1623.
12. Olkin I: **Odds ratios revisited** *Evidence-Based Medicine* 1998, **3:**71.
13. Senn S: **Odds ratios revisited** *Evidence-Based Medicine* 1998, **3:**71.
14. Sackett DL, Deeks JJ and Altman DG: **Down with odds ratios!** *Evidence-Based Medicine* 1996:164-166.
15. Baker SG and Lindeman KS: **The paired availability design: A proposal for evaluating epidural analgesia during labor** *Statistics in Medicine* 1994, **13:**2269-2278.
16. Angrist JD, Imbens GW and Rubin DR: **Identification of causal effects using instrumental variables** *Journal of the American Statistical Association* 1996, **92:**444-455.
17. Baker SG and Lindeman KS: **Rethinking historical controls** *Biostatistics* 2001, **2:**383-396.
18. Baker SG, Lindeman KS and Kramer BS: **The paired availability design for historical controls** *BMC Medical Research Methodology* 2001, **1:**9[http://www.biomedcentral.com/1471-2288/1/9].
19. Cuzick J, Edward R and Segnan N: **Adjusting for non-compliance and contamination in randomized clinical trials** *Statistics in Medicine* 1997, **16:**1017-1029.
20. Baker SG: **The paired availability design: an update** *In: Nonrandomized Comparative Clinical Studies* Edited by: *Abel U, Koch A. Dusseldorf: Medinform-Verlag*; 1998:79-84.
21. Deeks JJ: **Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes.** *Statistics in medicine* 2002:1575-1600.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/3/10/prepub